# Cooperation in Human-Robot Interaction: Playing Prisoner's Dilemma with ChatGPT

Samuel Harrison & Laura Johnston

*Department of Statistical Science, University College London*

**Abstract**

This study investigates how cooperation in human-robot interaction is affected by the anthropomorphisation of computer partners, using the Iterated Prisoner's Dilemma game. The study builds upon research that suggests that anthropomorphising robots can build trust in humans. Participants played the game with a computer partner, and their level of cooperation was recorded. The study tested the effect of anthropomorphisation on human cooperation levels by randomly assigning partcipants' computer partners that had been anthropomorphised to different levels (no anthropomorphisation, text-based dialogue, and spoken dialogue). The computer partners also played different first moves (cooperate or betray) during the games.

The computer dialogue was generated using ChatGPT and included empathy-evoking language and humor to increase trust and cooperation. At the end of the game, participants were asked if they perceived their computer partner to be human-like.

The results of the study indicate that neither level of anthropomorphisation of the computer partner, nor their first move, had a significant effect on the level of cooperation observed between the human participant and computer partner. However, participants' perception of human-like qualities in their computer partner may have impacted the level of cooperation.

## 1. Introduction

Human-robot interaction research aims to promote cooperation and trust between humans and robots. Attributing human characteristics to non-human entities, such as giving dialogue and voice to computers and building robots with faces and bodies, is defined as *anthropomorphisation*. Cominelli et al. (2021) investigated the impact of anthropomorphisation on trust and found that anthropomorphising robots can build trust in humans, but only if the non-human partner is perceived to be human-like. This research aims to expand on these findings by investigating human cooperation with anthropomorphised computer partners.

Cooperation within social science research is often studied using the Prisoner's Dilemma (Kreps et al., 1982). The original game involves two players (prisoners) who can choose to cooperate or betray each other. Their decision impacts their "prison sentence", as well as their partners. Players must decide whether to prioritise their own individual gain (a shorter prison sentence) over mutual benefit (a shorter combined sentence for both). An example payoff matrix for the game can be found in Table 2.

Multiple rounds of the Prisoner's Dilemma are played in the Iterated Prisoner's Dilemma. This repeated version has been used to research social decision making (Axelrod, 1980) since the game allows cycles of cooperation and a trusting relationship to develop between players. In this research, the Iterated Prisoner's Dilemma game was used to measure cooperation between human and computer partners.

The primary aim of this research is to investigate how human cooperation is impacted by the anthropomorphisation of computer partners. This intends to contribute to the understanding of whether humans are more likely to cooperate with computers or robots that are perceived to be human-like.

## 2. Design

### 2.1. The experiment

To achieve this objective, participants played the Iterated Prisoner's Dilemma game with a computer partner. Ten rounds ensured participants stayed engaged in the game, whilst allowing enough time for a relationship to be formed between partners. Each round, participants had to choose whether to cooperate or betray their computer partner. Participants could see the computer's choice between each round.

The level of anthropomorphisation applied to the computer partner was the primary treatment of interest. Computer partners had three possible levels of anthropomorphisation; A0: no anthropomorphisation; A1: text-based that displayed dialogue in written form; and A2: voice-based with spoken dialogue. Levels A1 and A2 used the same dialogue, which was communicated to the human player in-between rounds.

The computer dialogue was intended to be authentic and human-like. Dialogue features known to increase trust and cooperation were included, such as empathy-evoking language (Bickmore, 2001) and humour (Jung, 2003). All computer dialogue was generated by the ChatGPT (Table 1). Dialogue was then recorded for the spoken level A2.

| Game scenario | Computer dialogue example |
|---|---|
| Cooperation cycle (both the human and computer have been cooperating with one another) | *Let's stick together like glue, we'll be able to get a reduced sentence if we both remain silent.* |
| Computer retaliation (the computer has just betrayed the human) | *I know I made a mistake by betraying you last round, but I was just trying to look out for myself. I understand if you're feeling hurt right now, but I hope you can find it in your heart to give me another chance. Together, we can make things right.* |

Table 1: Two examples of computer dialogue for different game scenarios. To create the computer dialogue, a bank of 32 phrases was generated by ChatGPT by prompting a specific style of language and a given scenario in the game. The generated dialogue in the first example uses humour, while the second example uses empathy-evoking language.

A tit-for-tat strategy was found to encourage cooperation in the Iterated Prisoner's Dilemma (Oskamp, 1971). Therefore, the computer partner was programmed to play a tit-for-tat strategy. As this mirrored the human player's previous move, the first move of the computer was also instructed. Therefore, the first move made by the computer partner was introduced as a secondary treatment; F0: cooperate and F1: betray. This gave six combined treatment levels in total.

Once the game had concluded, participants were asked *"did you perceive your computer partner to be human-like?"* Answers to this question were recorded alongside the response variable.

## 2.2. *Response variable*

The combined payoffs for each round reflect the level of cooperation between the participant and computer partner. These values can be seen in Table 2. Totaling the combined payoffs for all ten rounds of the game gave the response variable for the experiment. Since the total combined payoff, $Y$, quantified the level of cooperation for the game, $Y$ was an appropriate choice of response variable.

Furthermore, $Y$ ranged from 0 to 100. A response of 0 represented no cooperation between the participant and computer for all ten rounds, whereas a response of 100 represented full cooperation. This was intentional to be easily interpreted.

| | | *Computer* | |
|---|---|---|---|
| | | *Cooperate* | *Betray* |
| *Human* | *Cooperate* | 5, 5 (**10**) | −2, 8 (**6**) |
| | *Betray* | 8, −2 (**6**) | 0, 0 (**0**) |

Table 2: Payoff matrix containing payoff values for each player in one round of the Prisoner's Dilemma. Combined payoffs, given in bold, are high when both players cooperate, moderate when only one player cooperates, and low when neither player cooperates.

## 2.3. *Pilot study*

A pilot of the experiment was completed by five participants. All participants were assigned the highest anthropomorphisation level, A2, with the computer partner choosing to cooperate on the first move, F0. The results gave an estimation of the distribution of the response variable for one treatment combination.

Feedback was also collected from the pilot subjects about their experience, enabling the identification of program defects and highlighted overlooked confounding variables. For example, the user interface displayed a round counter. However, it was suggested this influenced the decision making of players by guessing when the game would end. The counter was therefore removed.

## 2.4. *Sample size*

The number of replications per treatment combination, $m$, was estimated using the equation for the 95% confidence interval for the difference between group means. In the experiment, a difference in response greater than 15 was deemed significant, since this is equivalent to a difference of more than one human cooperation. Therefore, the size of confidence interval for the difference of group means, $L$, needed to be less than 30. This is given by the equation

$$L = 2t_{0.025}(6(m-1))s\sqrt{\frac{1}{m}} < 30$$

where $s$ is the standard error of the residuals.

80% of the scores from the pilot were within the interval $67.2 \pm 19.2$, where 67.2 was the mean. If the distribution was assumed to be normal, this interval would correspond to $2.56s$ in length. Therefore, $s = 15.4$. The smallest integer to satisfy the inequality is $m = 5$. Hence, a total sample size of 30 was required for the experiment.

### 2.5. Confounding variables

#### 2.5.1. User interface design

The user interface was designed to control for confounding factors. A simple click-based user interface was created using Python and HTML programming, enabling participants to make decisions easily. The game's introduction was carefully written not to influence participants' perspective of the game. A computer voice with a neutral tone and accent was selected to eliminate any potential biases, alongside being easily understood by all participants.

#### 2.5.2. Participants

Thirty MSc Data Science students from University College London volunteered for the experiment. Participants were 21-29 years old with an equal gender split and experienced with programming and ChatGPT. Age, gender, and technological affinity were therefore controlled for.

#### 2.5.3. Environment

The experiment was conducted over two weeks. Participants played the game in a private environment and wore headphones to minimise distractions.

### 2.6. Randomisation

Participants were randomly assigned a computer partner programmed with one of the six treatment combinations.

### 2.7. Modelling

The data was analysed using a linear model with corner-point parametrisation. The computer partner with no anthropomorphisation (A0) that cooperated on the first move (F0) was chosen as the reference level to compare against. The model coefficients of interest therefore measured the effect of adding anthropomorphisation.

The model used numeric indicator variables for the anthropomorphisation and first move levels. Interaction terms were also included, allowing any dependencies between variables to be recognised. The resulting linear model was in the form:

$$\begin{aligned} Y = \beta_{A0,F0} &+ \beta_{A1}x_{A1} + \beta_{A2}x_{A2} + \beta_{F1}x_{F1} \\ &+ \beta_{A1,F1}x_{A1}x_{F1} + \beta_{A2,F1}x_{A2}x_{F1} + \epsilon \end{aligned}$$

where

$\beta_{A0,F0}$ is the baseline treatment level,

$x_{A1}$, $x_{A2}$, $x_{F1}$ are indicator variables for the treatments,

The other $\beta$ coefficients measure the effect of applying the different treatments,

$\epsilon \sim N(0, \sigma^2)$.

A linear model was chosen to model the data as there were two treatment variables. Furthermore, linear models are flexible to add or remove covariates, allowing adjustment to control for confounding variables.

## 3. Data and Analysis

### 3.1. The data

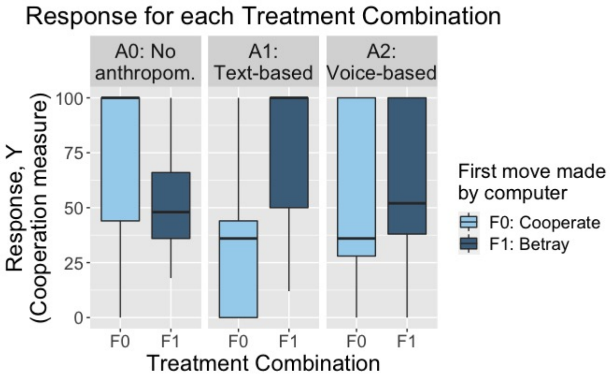

Figure 1: A boxplot to show the distribution of the response variable, $Y$, for each of the six treatment combinations (three levels of anthropomorphisation; A0, A1, A2; with two levels of first move; F0, F1)

A boxplot for all participant responses for each treatment combination is displayed in Figure 1. There does not appear to be any clear trend between the anthropomorphisation treatments levels, suggesting that anthropomorphisation did not impact human cooperation.

### 3.2. Analysis

The linear model assumes normally distributed error terms with constant variance. Diagnostic plots (Figure 2) support the assumption of constant variance and suggest normal errors with some deviation. The Shapiro-Wilk test (Shapiro and Wilk, 1965) confirmed no significant departure from normality at the 0.05 level ($p = 0.14$).

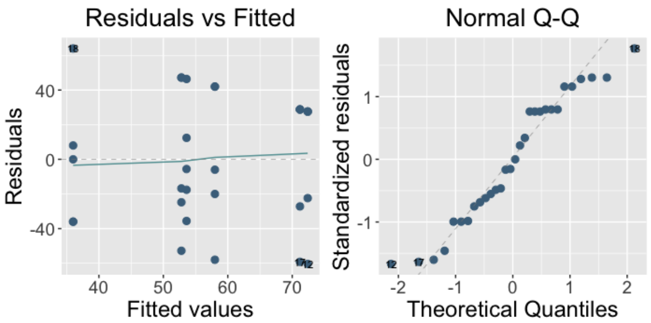

Figure 2: Two diagnostic plots for the fitted linear model. Points in the QQ-plot fall along the straight $y = x$ line when residuals are normally distributed. Points in the Residuals vs Fitted plot are evenly spread either side of the horizontal line when residuals have a constant variance.

To test the significance of the fitted model, an F-test was used to compare the model against the null model, H0: $\beta_{A1} = \beta_{A2} = \beta_{F1} = \beta_{A1,F1} = \beta_{A2,F1} = 0$. This test compared the variability of the fitted model with the variability of the model without any treatment variables (the null model). The test statistic followed the F(5, 24) distribution under the null model, H0. The null model could not be rejected at the 0.05 level ($p$ = 0.73).

## 4. Discussion

Since the null model could not be rejected, none of the treatment coefficients were significantly different to zero. Therefore, there is not enough evidence to determine if anthropomorphisation or the computer's first move explained human cooperation. This might be due to the experiment's limitations.

Moreover, in the study by Cominelli et al. (2021), trust in non-human partners was impacted by the perceived degree of human-likeness of the partner. Therefore in this experiment, to test whether participants perception of their computer partner impacted cooperation, participants answer to the post-experiment question *"did you perceive your computer partner to be human-like?"* were added to the model as a binary covariate, $x_H$. Performing backward stepwise regression on this model resulted in a linear model including only the $x_H$ covariate. An F-test on this model rejected the null model ($p$ = 0.0072), which suggests perception of the computer partner's human-likeness is related to cooperation.

However, it should be emphasised that the $x_H$ covariate was not designed to be an experimental variable. Thus, inference should not be made about the exact relationship between perception and cooperation, since other variables may have affected participants' answers to the question. For example, participants were aware of the final game outcome before responding to the question, which may have affected their willingness to acknowledge human-like qualities of the computer partner.

Future studies should aim to explore, in greater depth, the impact of anthropomorphisation and the role of perception in determining human cooperation with computer partners. Future work should also aim to address the limitations in this experiment outlined below.

### *4.1. Limitations*

#### *4.1.1. Levels of anthropomorphisation*

The levels of anthropomorphisation were intended to influence participants' perception of human-like qualities in their computer partner. However, the data shows 40% of participants perceived their computer partner to be human-like at both A0 and A1 levels, increasing to 60% at A2. Given the small jump, this might suggest that the levels of anthropomorphisation were not sufficiently distinct from the reference level.

#### *4.1.2. Sample size*

The standard error of the pilot study was estimated to be 16.6. However, the experiment's standard error was 40.5. This necessitates a larger total sample size of 174 to attain the required precision.

#### *4.1.3. Participants*

Prior knowledge of the Prisoner's Dilemma influenced many participants to adopt a strategy-focused approach to the game. Therefore, the sample was not necessarily representative of the wider population.

### *4.2. Conclusion*

This study investigated the impact of anthropomorphising computer partners on human cooperation in the Iterated Prisoner's Dilemma game. The results suggest that anthropomorphisation did not significantly impact the level of cooperation between human and computer partners. However, indistinct levels of anthropomorphisation, a small sample size, and participant's prior knowledge could have limited the experiments outcomes. These findings contribute to the ongoing research on promoting cooperation during human-robot interactions.

## 5. References

Axelrod, R. (1980). Effective Choice in the Prisoner's Dilemma. The Journal of Conflict Resolution, 24(1), 3–25. http://www.jstor.org/stable/173932

Bickmore, T., & Cassell, J. (2001). Relational agents. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/365024.365304

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., Nardelli, M., Scilingo, E. P., & Kirchkamp, O. (2021). Promises and trust in human-robot interaction. Scientific Reports (Nature Publisher Group), 11(1), 9687–9687. https://doi.org/10.1038/s41598-021-88622-9

Jung, W. E. (2003). The Inner Eye Theory of Laughter: Mindreader Signals Cooperator Value. Evolutionary Psychology, 1(1), 147470490300100. https://doi.org/10.1177/147470490300100118

Kreps, D.M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners'dilemma. Journal of Economic Theory, 27(2), 245–252. https://doi.org/10.1016/0022-0531(82)90029-1

Oskamp, S. (1971). Effects of Programmed Strategies on Cooperation in the Prisoner's Dilemma and Other Mixed-Motive Games. The Journal of Conflict Resolution, 15(2), 225–259. http://www.jstor.org/stable/173471

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591–611. https://doi.org/10.2307/2333709

Word count: 1976